We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Towards conformal interstitial light therapies: Modelling parameters, dose definitions and computational implementation

Emma Henderson, William C. Y. Lo and Lothar Lilge
Department of Medical Biophysics, University of Toronto
Canada

1. Introduction

External beam radiation therapy and high dose-rate brachytherapy were among the first tested medical applications for simulated annealing as an inverse planning optimization algorithm. The choice of these two applications is justified, as the tissue response to a given radiation dose is well established across different tissue types, leading to standardized clinical dosimetry concepts. A brief overview of the current state of simulated annealing in radiotherapy treatment planning is provided in Section 2. By contrast, the use of this popular optimization technique for interstitial light therapies, such as interstitial photodynamic therapy (IPDT), interstitial laser hyperthermia (ILH) and interstitial laser photocoagulation (ILP), requires an appreciation of the unique dosimetry challenges and the resulting development of new computational tools adapted to the requirements of light transport in tissue. Light transport is governed by interaction coefficients varying as a function of wavelength and differs between tissue types and between individuals.

This chapter introduces the clinical motivation behind conformal interstitial light therapies, the concept of using light dosimetry as a basic approach to clinical dosimetry, as well as our ongoing work on creating a computational framework for real-time light and clinical dosimetry (***http://code.google.com/p/gpu3d/***) , in order to extend the benefit of simulated annealing. This presents a very timely application for simulated annealing, as these light-based therapies carry the possibility of low-cost general treatments in oncology and an alternative to surgical intervention. Presently, the efficacy of light-based therapies is, to a certain extent, limited by currently used heuristic treatment plans to illuminate the clinical target volume with sufficient energy or power density. Individualized treatment plans, tailored to a specific patient's anatomy and preferably also the local light-tissue interaction coefficients, can overcome the shortcomings of the heuristic treatment plans, and thus maximize the benefits and impact of light-based treatments.

2. Treatment planning using simulated annealing

To help readers understand how simulated annealing can be applied clinically for treatment planning, this section reviews the progress made in the well-established field of radiotherapy treatment planning. This provides the basis for the discussion of emerging interstitial light

therapies and the clinical utility of simulated annealing in planning for these novel treatment options.

2.1 Simulated annealing in radiotherapy

Simulated annealing has been used as an optimization technique for radiation treatment planning in the clinical setting, with successes reported in both external beam radiation therapy (EBRT) (Aubry et al., 2006; Beaulieu et al., 2004; Morrill et al., 1995) and high dose-rate (HDR) brachytherapy (Lessard & Pouliot, 2001; Martin et al., 2007). For EBRT, the basic optimization problem is the determination of the appropriate temporal and spatial arrangements of multiple external radiation beams, which can be tailored in highly sophisticated and precise ways in modern 3-D conformal radiation therapy. For HDR-brachytherapy (which delivers a high radiation dose directly by implanting intense radioactive sources within the tumour for a short time), the optimization involves adjusting the duration (or *dwell time*) that a source pauses or *dwells* at each position (called *dwell position*) along the implanted catheter. Treatment delivery is accomplished using a computer-controlled robotic unit called a *stepping source device* or an *afterloader*, which moves the radioactive sources (commonly ^{192}Ir) along the catheters according to the optimized dwell time distribution in order to deliver the desired radiation dose distribution.

There is now a shift in paradigm as we strive to achieve the century-old objective of delivering a curative radiation dose to the tumour while sparing sensitive structures and surrounding normal tissues. That is, instead of manually specifying the treatment parameters and repeatedly evaluating the resulting radiation dose distribution (*forward planning*), a desired dose distribution is prescribed by the physician and the task of finding the appropriate treatment parameters is automated with an optimization algorithm (*inverse planning*). The latter approach, or inverse planning, is much more goal-oriented and efficient.

The concept of inverse planning, using simulated annealing as the optimization engine, can be briefly summarized as follows. Note that the following description focuses on HDR-brachytherapy and is greatly simplified, although the steps are quite similar for EBRT.

1. **3-D imaging of anatomical structures:** The first step typically involves the definition and contouring of the anatomical structures of interest using 3-D imaging modalities such as ultrasound, computed tomography (CT) and magnetic resonance imaging (MRI). Functional imaging techniques such as magnetic resonance spectroscopy (MRS) and positron emission tomography (PET) are also investigated for better localization and targeting of the tumour, thus permitting the administration of an escalated dose to the target (Scheidler et al., 1999). The anatomical structures of interest include the entire tumour and an estimated margin for the microscopic tumour spread, together referred to as the *clinical target volume* (CTV), as well as surrounding normal tissue and *organs at risk* (OAR). For convenience, all these structures are considered the *treatment planning volume* (TPV) in this chapter.
2. **Specification of desired dose distribution:** For each anatomical structure or organ, the desired 3-D dose distribution is prescribed by the physician through a set of dose constraints. Using the prostate as an example, a physician may define a minimum acceptable radiation dose D^{min} to the prostate (CTV) to ensure complete coverage and a maximum permissible dose D^{max} to the surrounding structures such as the rectum or urethra (OAR) to avoid complications. Each structure can be assigned a different set of dose constraints (D^{min} and D^{max}). A weighting factor w is also used to specify the relative importance of meeting a specific set of dose constraints or clinical objectives. A

penalty value p_i, computed using Equation 1 (Lessard & Pouliot, 2001), can be assigned to each dose point i (usually chosen to be either on the surface or inside the volume of the CTV or OAR) after calculating the dose distribution D_i resulting from a proposed set of treatment parameters. For details on brachytherapy dose calculations, see the American Association of Physicists in Medicine (AAPM) Task Group No. 43 Report (Rivard et al., 2004).

$$p_i = \begin{cases} w^{min}(D_i - D^{min}) & \text{if } D_i < D^{min} \\ w^{max}(D_i - D^{max}) & \text{if } D_i > D^{max} \\ 0 & \text{if } D^{min} <= D_i <= D^{max} \end{cases} \quad (1)$$

To evaluate a treatment plan, a global penalty function called a *cost function* or *objective function*, denoted E here, can be defined by summing the penalty values over all dose points for every anatomical structure. Note that variations of the cost function defined above exist, but the basic idea is similar in most cases. That is, the closer the calculated dose distribution is to the prescribed distribution, the lower the value of the cost function becomes.

3. **Optimization of treatment parameters:** The objective becomes minimizing the cost function, E, by solving for the appropriate combination of treatment parameters required to meet the clinical objectives. For example, in HDR-brachytherapy, one of the treatment parameters that needs to be optimized is the dwell time values (or dwell time distribution) for the different source positions or radioactive seed positions (also known as dwell positions) to achieve the desired dose coverage (Lessard & Pouliot, 2001). To illustrate how simulated annealing can be applied to treatment planning in HDR-brachytherapy, a generic pseudo code is shown in Algorithm 1 as a simplified example. (Note that the exact implementation details, such as the choice of the terminating condition, the definition of the cost function, or the selection of the cooling schedule, can differ depending on the clinical scenario and computational resources available.) First, the initial dwell time values are set and the corresponding cost function E_0 for this initial distribution is computed. Then, the dwell time value for a randomly chosen dwell position is randomly incremented or decremented. This change leads to a new dwell time distribution, from which a new dose distribution D_i and its associated penalty p_i can be computed. Next, the global penalty value or cost function E_k is compared against the previous one E_{k-1}. If the new dwell time distribution leads to a lower cost function ($\Delta E < 0$), then the change is accepted. Otherwise, the new treatment parameter is accepted with a probability of $P(\Delta E) = exp[-\Delta E/T(k)]$. For shorter computation time, a faster cooling schedule (fast simulated annealing) can be used by defining $T(k) = T_0/k^{\alpha}$ where T_0 is the initial temperature and α is the speed parameter. The entire process can be repeated until the cost function has reached a threshold as defined by clinical requirements or when further iterations do not produce a clinically significant difference.

The end result of this inverse planning process is an optimized set of treatment parameters forming the individualized treatment plan - in this case, a dwell time distribution delivered by an afterloader that produces a clinically acceptable radiation dose distribution tailored to an individual patient's anatomy. The notion of precisely shaping the dose distribution to match one's anatomy forms the basis of conformal radiation therapy.

Algorithm 1 Example of using simulated annealing for treatment planning

Initialize
while $E > threshold$ **do**
  Modify treatment parameters
  Compute new dose distribution D_i //**NOTE: Dose definition and dose calculation for radiotherapy and light-based therapies are fundamentally different**.
  Assign penalty p_i (Eq. 1)
  Compute cost function $E_k \leftarrow \sum p_i$
  $\Delta E = E_k - E_{k-1}$
  $P(\Delta E) = exp[-\Delta E/T(k)]$
  if $\Delta E < 0$ **then**
    Accept new treatment parameters
  else
    Accept new treatment parameters with a probability of $P(\Delta E)$
  end if
  $k \leftarrow k + 1$
end while

2.2 Fundamental differences between radiation therapy and light-based therapies

As the mass attenuation coefficient of ionizing radiation is on the order of < 0.3 cm^2/g for clinical applications and scattering of high-energy photons is weak, conformality in radiation therapy - that is, achieving high dose within the CTV combined with a steep dose gradient at its boundary and resulting low dose in surrounding tissue and OAR - can be achieved by superimposing radiation fields emanating from implanted sources, as described above, or from external beams so that they overlap at the CTV. For external beams, further conformality can be achieved by spatially modifying the exposure over each beam's cross section, as in intensity-modulated radiation therapy (IMRT) (Bortfeld, 2006). However, in light-based therapies, the very high attenuation coefficient and scattering coefficient necessitate different means to achieve conformality. For interstitial PDT in particular, the parameter space is given by the number and emission properties of implantable optical fibres and the length over which they emit. While for ionizing radiation the tissue interaction coefficients of the CTV and the OAR do not vary appreciably, light can encounter rather large differences in its interaction coefficients with tissue, ranging from, for example, < 2 cm^{-1} to 30 cm^{-1} in pig muscle and rabbit liver, respectively. Achieving conformality is further complicated if the other efficacy-determining parameters are not homogeneously distributed across the TPV, as discussed next. Despite these fundamental differences between radiation therapy and light-based therapies, the optimization algorithm based on simulated annealing (shown in Algorithm 1) can be similarly applied for planning light-based therapies, although the dose computation step would differ significantly as new definitions of dose are required for these novel therapies.

3. Dosimetry concepts for light-based therapies

This section provides the clinical background for interstitial light therapies as well as their unique treatment planning problem.

3.1 Dose definitions

For any medical therapy, the preferred definition of "dose" is a measurable quantity directly correlated with the desired biological or clinical outcome. In radiotherapy, this quantity is the energy absorbed by the tissue, shown to correlate with the tissue's response, and is possibly modulated by various external and internal factors such as tissue hypoxia. While the sensitivity of different tissues varies somewhat, the threshold to induce cell death varies by less than a factor of 2. Because the attenuation of ionizing radiation in soft tissues is low and is a function of its quantum energy, the energy density to be delivered can be directly calculated. The dose definition is less clear for ILH and ILP due to the biological response to temperature increase and heat transfer: active, convection through the vascular system, or passive, by diffusion. These two parameters depend on the extent of the vascular system, tissue density, and other structural tissue properties. For photodynamic therapy (PDT), owing to its complex mechanism requiring a drug and molecular oxygen, different efficacy-determining parameters apply. The dose definitions for these three therapies are further discussed below.

3.1.1 Interstitial Laser Photocoagulation (ILP) and Interstitial Laser Hyperthermia (ILH)

When describing photothermal applications such as ILP and ILH, the Bioheat Equation (Pennes, 1948) needs to be considered:

$$\rho_t c_t \frac{\partial T(\vec{r},t)}{\partial t} = \nabla(k_t \nabla T(\vec{r},t)) - \omega_b c_b \rho_b (T_{art}(\vec{r},t) - T(\vec{r},t)) + S(\vec{r},t) \tag{2}$$

In Equation 2, $S(\vec{r},t)$ describes the heat source distribution as a function of space and time. The resulting thermal energy distribution in the tissue, $\partial T(\vec{r},t)/\partial t$, is affected by the heat capacity of the vascular system c_b and tissue c_t, the thermal conductivity of the tissue k_t, as well as the general heat diffusion or blood perfusion rate ω_b throughout the tissue. Other important quantities include the density ρ and the temperature of the arterial blood $T_{art}(\vec{r},t)$. Note that additional terms can be added to Equation 2 to account for other effects such as metabolic heat generation and evaporation of water from the tissue. Equation 2 is sufficient for long light exposure when a steady state of the heat distribution is achieved. For short exposure time, and when using a pulsed light source, the actual damage for a given temperature is better described by the Arrhenius Integral (Henriques Jr & Moritz, 1947), given in Equation 3, which also considers the tissue's activation energy E_α, and tissue specific factors, A, temperature, T, and universal gas constant, R. The thermal damage parameter, denoted $\Omega(\vec{r},t)$ [dimensionless], is a measure of the degree of thermal injury and is computed as follows:

$$\Omega(\vec{r},t) = \int_0^t A e^{\frac{-E_\alpha}{RT(\vec{r},t)}} dt \tag{3}$$

Hence, accurate dosimetry for laser thermal therapies requires the knowledge of both the optical properties (absorption coefficient μ_a and scattering coefficient μ_s) and thermal properties (E_α) of tissue. Another significant complication is the temporal variation of these optical and thermal parameters, such as through the thermal transition from normal to, for example, coagulated tissue. Hence, simulated annealing based optimizations would need to be executed in a temporally resolved manner to account for such temporal variations in these thermal therapies.

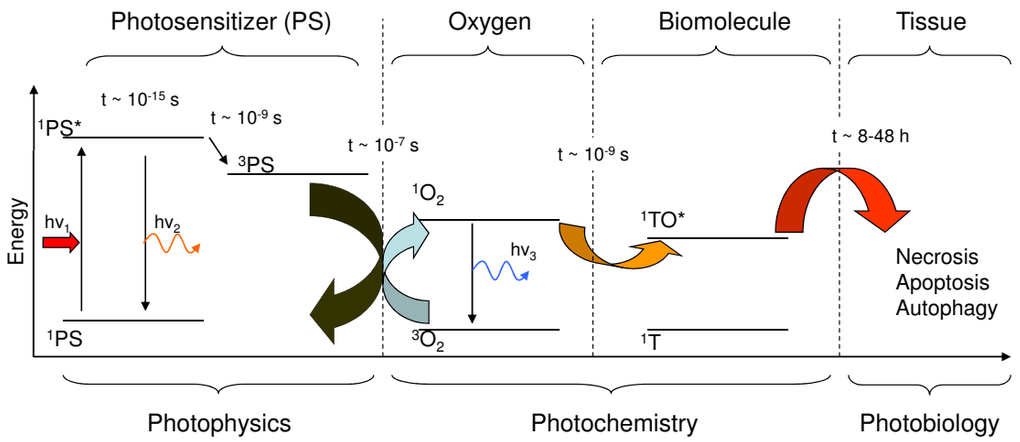


Fig. 1. The photophysics, photochemistry, and photobiology of PDT. The symbol $h\nu_1$ represents the incoming light, $h\nu_2$ represents the fluorescence by the photosensitizer (PS), and $h\nu_3$ is represents the singlet oxygen phosphorescence.

3.1.2 Interstitial Photodynamic Therapy (IPDT)

Before settling on a definition of dose for IPDT, one must understand the mechanism of its therapeutic effect. Figure 1 shows the photophysical, photochemical, and photobiological steps leading from the initial light absorption to cell apoptosis or tissue necrosis. The quantum energy of a photon is absorbed by the photosensitizer (PS), lifting it into an electronic excited singlet state. Photosensitizers have a high intersystem quantum yield, trapping them for microseconds in the triplet state. Collisional exchange of spin and energy with ground state molecular oxygen ($^3O_2$), which is a triplet, can result in highly reactive singlet oxygen ($^1O_2$) and ground state PS. Singlet oxygen has a lifetime in the low nanoseconds in biological systems and oxidizes lipids and proteins within its range indiscriminately ($^1T \rightarrow {}^1TO^*$). Accumulation of oxidative damage disrupts normal cell function and triggers apoptosis or necrosis. From this description, it should be clear that

1. Singlet oxygen is the main cytotoxic mediator of PDT-induced damage (Weishaupt et al., 1976);
2. There are three main parameters that govern $^1O_2$ production: light energy density, concentration and molar extinction coefficient of PS, and concentration of $^3O_2$. Note that all three parameters must be spatially and temporally co-localized.

Although $^1O_2$ fulfills the requirement for an ideal dose metric by virtue of directly correlating with the biological outcome, its quantification *in vivo* is very difficult and not currently feasible for the desired interstitial applications. Its characteristic phosphorescence at 1270 nm has a low quantum yield and there are few detectors with sufficient sensitivity at this wavelength. The strategy, then, is to deduce $^1O_2$ deposition based on the other PDT efficacy determining parameters (i.e., light, PS, and oxygen). This approach is called *explicit* or *extrinsic* since $[^1O_2]$ is directly calculated based on the interactions of its precursors (Niedre et al., 2003). By contrast, the *implicit* or *intrinsic* approach quantifies an interim photoproduct, the dynamics of which is directly related to $[^1O_2]$ and thus serves as a surrogate which *implies* the production of $^1O_2$. A possible interim photoproduct is the excited singlet state of the PS, quantified through its fluorescence intensity (Pogue, 1994).

Among the explicit approaches to PDT dosimetry, one will find both empirical and analytical models which aim to quantify the tissue response for a set of conditions. Empirical models include the critical fluence model (Jankun et al., 2004) and the threshold model (Farrell et al., 1998).

The critical fluence model defines the dose in terms of the light parameter alone; the quantity of interest is the total fluence (in units of joules per unit area) delivered to the target. Details on how this calculation is carried out will be explained in Section 3.2. When employing this approach, one must assume that PS and $^3O_2$ are present in unlimited quantities throughout the target for the duration of irradiation.
In the threshold model, the dose definition is the number of photons absorbed by the photosensitizer per unit mass of tissue. Here, only ubiquitous molecular oxygen availability is assumed. Information about the light distribution within and surrounding the target is combined with information about the PS distribution, considering tissue specific responsivity.
Two analytical approaches include the $^1O_2$ production (Zhu et al., 2007) and oxygen consumption models (Wang et al., 2007), which solve a set of differential equations representing the photophysical and photochemical reactions in PDT, assuming constant interaction coefficients.
Each of the models discussed above employs different definitions of dose and makes different assumptions about the efficacy-determining parameters. The utility of each will depend on the particular clinical problem. For clinical targets that are well-vascularized and confined to surfaces (some examples include the skin, oesophagus, or bladder), one may confidently apply the critical fluence or threshold model as the assumptions are likely justified. When the target is a solid tumour in an enclosed site, such as the prostate (Aniola et al., 2003), or a large, advanced tumour, these models are likely to fail. Solid tumours are known to have hypoxic regions, and drug distribution within them is often heterogeneous (Di Paolo & Bocci, 2007). Furthermore, these deep-seated tumours require an interstitial approach to light delivery which will render treatment planning even more difficult. In spite of these known issues, the prescribed dose used in the clinic remains the light energy density alone, due to limitations in the current technical ability to collect all required data in a spatially resolved manner. Typically, light energy density is calculated by applying the diffusion approximation to the transport equation along with finite-element based methods and using population-averaged tissue optical properties. These techniques are further detailed in Section 3.2. Optimizing conformal IPDT thus becomes an iterative process to attain the required minimum "dose" across, preferably, the entire CTV without exceeding the maximum permissible dose for the OAR. Minimization is executed in an N-dimensional parameter space comprising nominal (number of optical fibres) and interval (length, orientation, emission profile, and total power delivered for each fibre) data.

3.2 Light dosimetry models

With the exception of the oxygen consumption model in PDT, all other dose metrics require knowledge of the light distribution or, more precisely, its propagation in tissue. Unlike ionizing radiation, two interaction coefficients - light scattering and absorption - need to be considered and their values are large (Cheong et al., 1990), leading to rather steep gradients of light energy density. Additionally, there is a large difference in tissue optical properties between organs and individuals, as is evident from the visible appearance of these organs. For most

wavelengths of interest in IPDT, ILH, and ILP, the light scattering coefficient is larger than the absorption coefficient and light transport follows the Boltzmann Transport equation. When distal to boundaries and light sources, light transport can be well approximated by diffusion theory.

3.2.1 Transport theory

Light has both wave-like and particle-like behaviours, and both treatments of it are widely accepted among physical scientists. In tissue optics, it is often more useful to consider light in terms of its particle-like properties owing to the heterogeneity in the coefficients of permittivity and permeability (ϵ_r, μ_r) within tissue. The quantities of interest, then, include

1. **Photon distribution**, $N(\vec{r},\hat{s},t)$ with units $[\frac{1}{m^3 sr}]$, representing the number of photons per unit volume propagating in the direction denoted $\hat{s}$ within solid angle $d\omega$ at position $\vec{r}$ at time t.
2. **Radiance**, $L(\vec{r},\hat{s},t) = h\nu cN(\vec{r},\hat{s},t)$ with units $[\frac{W}{m^2 sr}]$, where h is Planck's constant, ν is the frequency of light, and c is the speed of light. Note that the unit of radiance is the power per unit area per steradian.
3. **Fluence rate**, $\phi(\vec{r},t) = \int\limits_{4\pi} L(\vec{r},\hat{s},t)d\omega$ with units $[\frac{W}{m^2}]$, which is commonly used in tissue optics as it can be readily measured.

There are two quantities used to describe scattering: the scattering coefficient, μ_s, and the anisotropy, $g \equiv < cos\theta >$, where θ is the scattering angle (or deflection angle). μ_s is the probability of a scattering event per unit length, while g describes the average cosine of the scattering direction. Most tissues are forward-scattering and have g values of 0.9. These two terms are often combined into the reduced scattering coefficient, $\mu_s' \equiv (1-g)\mu_s$.
Absorption occurs when there is a match in energy between the incoming light and two electronic states of the chromophore upon which the light is incident. The absorption coefficient, μ_a, is the probability of an absorption event per unit length. It is spectrally dependent for a given chromophore, and for a given tissue may be represented as $\mu_a(\lambda) = \sum \epsilon_i(\lambda_i)C_i$, where ϵ_i and C_i are the extinction coefficient and concentration, respectively, for chromophore i in the tissue.
The radiative transport equation (RTE) (Ishimaru, 1977) is a description of photon transport through a medium, derived from conservation of energy:

$$\int\limits_V \frac{\partial N}{\partial t}\,dV = \int\limits_V q\,dV + \int\limits_V v\mu_s \int\limits_{4\pi} p(\hat{s}',\hat{s})N\,d\omega' dV - \oint\limits_S vN\hat{s}\hat{n}\,dS - \int\limits_V v\mu_s N\,dV - \int\limits_V v\mu_a N\,dV \quad (4)$$

The left-hand side of the equation represents the net change of photon distribution integrated over a small volume V. The first two terms on the right-hand side of Equation 4 include a source term (q = the number of photons emitted per unit volume, time, and steradian) and another term describing any photons that are scattered from direction $\hat{s}'$ into the direction of interest $\hat{s}$ (where $p(\hat{s}',\hat{s})$ is the scattering phase function and v is the speed of light in the medium), respectively. The three loss terms are, from left to right, those photons lost to boundary crossing (where S denotes the surface of the boundary and $\hat{n}$ is the unit normal pointing outwards), scattering out of the direction of interest, and absorption.

The above equation can be re-written, in terms of the radiance, and without integrating over volume:

$$\frac{1}{v}\frac{\partial L}{\partial t} = h\nu q + \mu_s \int_{4\pi} p(\hat{s}', \hat{s}) L \, d\omega' - \hat{s} \cdot \nabla L - \mu_s L - \mu_a L \tag{5}$$

There are only a few conditions for which an exact solution of Equation 5 is possible; therefore, simplification of the RTE is necessary. The first-order diffusion approximation, developed hereafter, is one such approach.

In this approach, the radiance, source term, and scattering function are expanded into a series of spherical harmonics; the first-order diffusion approximation truncates the series at the first-degree term. The radiance then becomes:

$$L(\vec{r}, \hat{s}, t) \approx \frac{1}{4\pi}\phi(\vec{r}, t) + \frac{3}{4\pi}\vec{F}(\vec{r}, t) \cdot \hat{s} \tag{6}$$

where $\phi(\vec{r}, t) = \int_{4\pi} L(\vec{r}, \hat{s}, t) d\omega$ is the fluence rate in units of $[W/m^2]$, while $\vec{F}(r, t) = \int_{4\pi} L(\vec{r}, \hat{s}, t)\hat{s} d\omega$ is the photon flux in units of $[W/m^2]$. The first term on the right hand side is isotropic, and the second is linearly anisotropic. Inserting Equation 6 into the RTE results in two coupled equations:

$$(\frac{1}{v}\frac{\partial}{\partial t} + \mu_a)\, \phi + \nabla \vec{F} = q_0 \tag{7}$$

$$(\frac{1}{v}\frac{\partial}{\partial t} + \mu_a + \mu_s')\, \vec{F} + \frac{1}{3}\nabla\phi = \vec{q_1} \tag{8}$$

Two assumptions are made at this point:

1. Sources are isotropic, i.e., the linearly anisotropic source term $\vec{q_1} = 0$.
2. Photon flux is in steady-state, i.e., $\frac{\partial \vec{F}}{\partial t} = 0$

We are left with Fick's Law:

$$\vec{F} = -\frac{1}{3(\mu_a + \mu_s')}\nabla\phi \tag{9}$$

with diffusion coefficient $D \equiv \frac{1}{3(\mu_a + \mu_s')}$. This is substituted into Equation 7 to obtain the Diffusion Equation:

$$\frac{1}{v}\frac{\partial \phi(\vec{r}, t)}{\partial t} - \nabla D(\vec{r})\nabla\phi(\vec{r}, t) + \mu_a(\vec{r})\phi(\vec{r}, t) = q_0(\vec{r}, t) \tag{10}$$

Since the assumption was made that sources are isotropic, diffusion theory may only be used when $\mu_s{}'$ is much larger than μ_a(a good rule of thumb is that $\mu_s{}' > 10\mu_a$) or when the point of interest is far (at least 1 mean free path, defined as $1/(\mu_a + \mu_s')$ (Jacques & Pogue, 2008)) from sources or boundaries. Small geometries are, therefore, excluded.

3.2.2 Finite-element method (FEM) based models

The finite-element method operates by first breaking the volume of interest into a mesh of discrete elements. The diffusion equation is then solved over these discrete elements, assuming a linear solution over the interpolation between nodes. FEM can handle heterogeneous tissue optical properties and complex geometries, depending on the size of discretization. The trade-off is the large amount of memory required, which will limit the mesh size, or number of nodes (Davidson et al., 2009).

3.2.3 Monte Carlo Method

The Monte Carlo (MC) method is a statistical sampling technique that has been widely applied to a number of important problems in medical biophysics and many other fields, ranging from photon beam modelling in radiation therapy treatment planning (Ma et al., 1999) to protein evolution simulations in biology (Pang et al., 2005). The name *Monte Carlo* is derived from the resort city in Monaco which is known for its casinos, among other attractions. As its name implies, one of the key features of the MC method is the exploitation of random chance or the generation of random numbers with a particular probability distribution to model the physical process in question (Metropolis & Ulam, 1949). Since the MC method inherently relies on repeated sampling to compute the quantity of interest, the development of the MC method has parallelled the evolution of modern electronic computers. In fact, initial interests in MC-based computations stemmed from von Neumann's vision of using the first electronic computer - the ENIAC (Goldstine & Goldstine, 1996) - for the modelling of neutron transport (Metropolis, 1989), which was later adopted for the development of the atomic bomb in World War II.

Despite the increased variety and sophistication of MC-based simulations today, most MC-based models still retain the same essential elements, including the extensive use of random numbers and repeated sampling. For example, in the case of photon transport, random numbers are used to determine the distance of photon propagation and the direction of scattering, among other interactions. Each photon is tracked for hundreds of iterations and typically thousands to millions of photons are required to accurately compute the quantity of interest, such as the light dose distribution. Due to the large number of iterations required, different variance reduction techniques (Kahn & Marshall, 1953) have been introduced to reduce the number of samples required to achieve a similar level of statistical uncertainty or variance in MC-based computations. Conversely, variation reduction schemes allow more equivalent samples to be computed within the same amount of time. Unfortunately, the simulation time remains high for solving complex optimization problems such as those for treatment planning, which require many of these MC simulations (as shown earlier in Algorithm 1). To circumvent this obstacle, we propose a novel computational framework to make the MC method a practical approach for light dosimetry in the next section.

4. Computational framework for light dosimetry

In biomedical optics, the MC method is considered the gold standard approach for modeling light transport in biological tissue due to its accuracy and flexibility in handling realistic 3-D geometry with heterogeneities in the light-tissue interaction coefficients. However, the use of MC simulations in iterative optimization problems, such as treatment planning for photodynamic therapy and other light-based therapies, has been hindered by its long computation time (Luu, J. et al., 2009; Lo, W.C.Y. and Redmond, K. et al., 2009; Lo, W.C.Y. et al., 2009). Hence, it is often replaced by diffusion theory, when homogeneous light-tissue interaction coefficients are assumed (Altschuler et al., 2005; Rendon, 2008), or diffusion theory in combination with the finite element method (Davidson et al., 2009; Johansson et al., 2007). Unfortunately, neither approach provides the flexibility desired for treatment planning. On the other hand, the iterative nature of treatment planning within an N-dimensional parameter space makes it infeasible for MC-based computation to become the core dose calculation method in simulated annealing. Overcoming this computational burden is essential for the clinical application of MC simulations and simulated annealing for treatment planning. Instead of using the traditional networked computer cluster approach, this section explores the

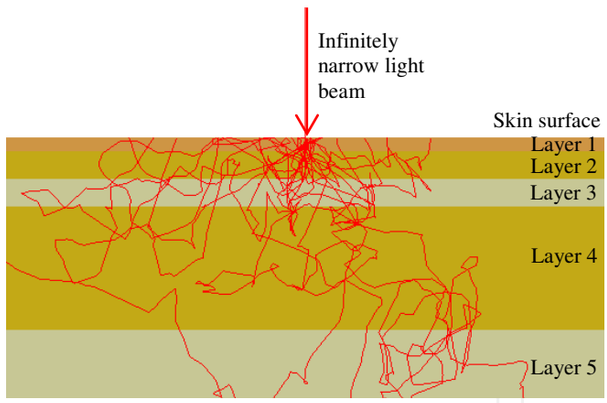


Fig. 2. MC simulation of photon propagation in a skin model (λ=633 nm).

use of modern computer graphics processing units (GPUs) for acceleration. To demonstrate the practicality of the GPU-based approach, a gold standard MC code package for modelling light propagation in multi-layered biological media (called MCML) was implemented on multiple NVIDIA GPUs. The final implementation was validated using an optical skin model to show the close correspondence between simulated isodose contours generated by the different computational platforms.

4.1 The MCML Algorithm

The MCML algorithm (Wang et al., 1995) models steady-state light transport in multi-layered turbid media using the MC method. The MCML implementation assumes infinitely wide layers, each of which is described by its thickness and its optical properties, comprising the absorption coefficient, scattering coefficient, anisotropy factor, and refractive index. A diagram illustrating the propagation of photon packets in a multi-layered skin geometry (Tuchin, 1997) is shown in Figure 2, using ASAP (Breault Research Organization, Tucson, AZ) as the MC simulation tool to trace the paths of photons (*ASAP - Getting Started Guide*, 2009).

In the MCML code, three physical quantities – absorption, reflectance, and transmittance – are calculated in a spatially-resolved manner. Absorption is recorded in a 2-D absorption array called $A[r][z]$, which represents the photon absorption probability density [cm^{-3}] as a function of radius r and depth z for a point source impinging on the tissue. Absorption probability density can be converted into more commonly used quantities in treatment planning such as photon fluence (measured in cm^{-2} for the impulse response of a point source). Fluence can be obtained by dividing the absorption probability density by the local absorption coefficient. To model finite-sized sources, the photon distribution obtained for the impulse response can be convolved with tools such as the CONV program (Wang et al., 1997).

The simulation of each photon packet consists of a repetitive sequence of computational steps and can be made independent of other photon packets by creating separate absorption arrays and decoupling random number generation for each group using different seeds. Therefore, a conventional software-based acceleration approach involves processing multiple photon packets simultaneously on multiple processors. Figure 3 shows a flow chart of the key steps in an MCML simulation, which includes photon initialization, position update, direction update, fluence update, and photon termination. Further details on each computational step may be found in the original papers by Wang et al.

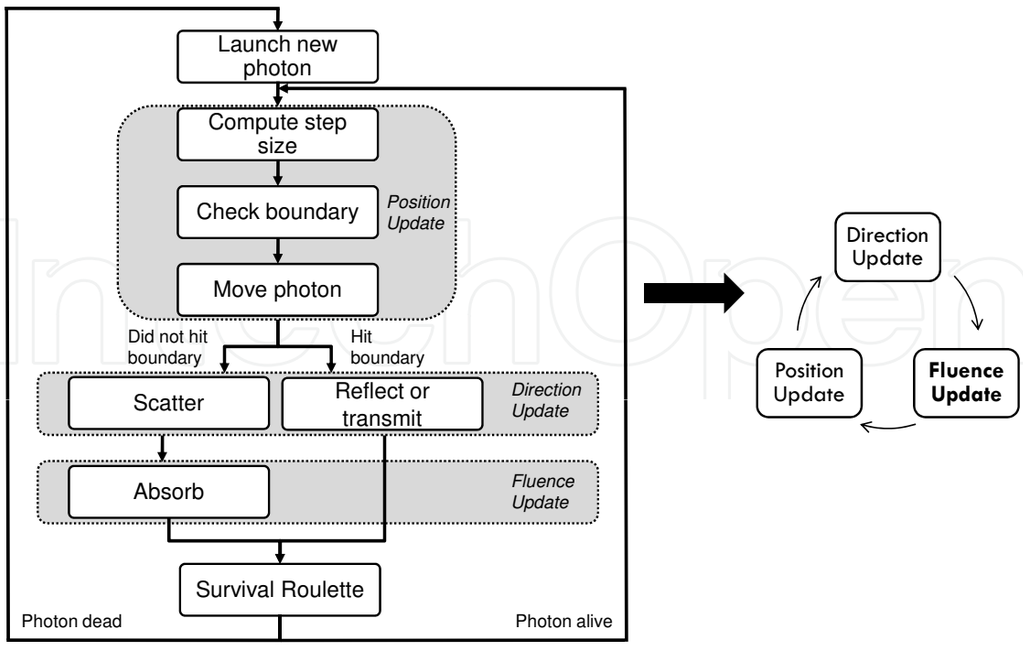


Fig. 3. Left: Flow-chart of the MCML algorithm. Right: Simplified representation used in subsequent sections.

4.2 Programming Graphics Processing Units with CUDA

The rapid evolution of GPUs and recent advances in general-purpose GPU computing have prompted the use of GPUs for accelerating scientific applications, including time-consuming MC simulations. This section introduces the key terminology for understanding graphics processing hardware, which was instrumental to the successful acceleration of the MCML code. Similarly, for other related applications, this learning curve is required to fully utilize this emerging scientific computing platform.

GPU-accelerated scientific computing is becoming increasingly popular with the release of an easier-to-use programming model and environment from NVIDIA (Santa Clara, CA), called CUDA, short for Compute Unified Device Architecture (*CUDA Programming Guide 2.3*, 2009). CUDA provides a C-like programming interface for NVIDIA GPUs and it suits general-purpose applications much better than traditional GPU programming languages. While some acceleration compared to the CPU is usually easily attainable, full performance optimization of a CUDA program requires careful consideration of the GPU architecture.

4.2.1 NVIDIA GPU Architecture

The underlying hardware architecture of a NVIDIA GPU is illustrated in Figure 4 (*CUDA Programming Guide 2.3*, 2009), showing both a unique processor layout and memory hierarchy. Using the NVIDIA GeForce GTX 280 GPU as an example, there are 30 *multiprocessors*, each with 8 *scalar processors* (SPs). Note that the 240 SPs (total) do not represent 240 independent processors; instead, they are 30 independent processors that can perform 8 similar computations at a time. From the programmer's perspective, computations are performed in parallel

by launching multiple *threads*, each containing a parallel unit of work. For example, a thread can simulate a group of photon packets in the MCML algorithm.

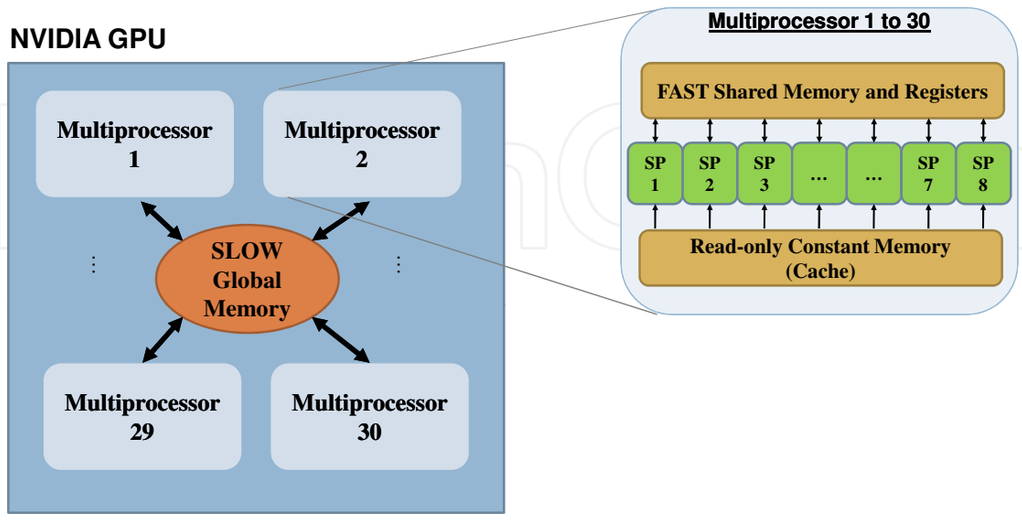


Fig. 4. Simplified representation of the NVIDIA GPU architecture for GTX 280

Second, the programmer must understand the different layers or types of memory on the GPU, due to the significant differences in memory access time. The outermost layer, which is also the largest and slowest (with a latency of up to 600 clock cycles), is the off-chip *device memory* (also known as *global memory*). Closer to the GPU are various kinds of fast, on-chip memories, including *registers* with typically a single clock cycle of access time, *shared memory* at close to register speed, and a similarly fast cache for *constant memory*. On-chip memories are roughly a hundred times faster than the off-chip memory; however, their storage space is limited. Finally, there is a region in device memory called *local memory* for storing large data structures, such as arrays, which cannot be mapped into registers by the compiler. As a result, it is important to map the computation efficiently to the different types of memories (e.g., depending on the frequency of memory accesses for different variables) to attain high performance.

4.2.2 Atomic Instructions

CUDA also provides a mechanism to synchronize the execution of threads using *atomic instructions*, which coordinate sequential access to a shared variable (such as the absorption array in the MCML code). Atomic instructions guarantee data consistency by allowing only one thread to update the shared variable at any time; however, in doing so, it stalls other threads that require access to the same variable. As a result, atomic instructions can give rise to performance bottlenecks. The cconcept of atomicity is illustrated in Figure 5.

4.2.3 Related Work

Previous attempts to use GPUs for MC-based photon simulations include the work by Alerstam et al., who reported ∼1000x speedup on the NVIDIA GeForce 8800GT graphics card compared to an Intel Pentium 4 processor. Their implementation simulates time-resolved photon migration (for photon time-of-flight spectroscopy) in a homogeneous, semi-infinite geometry (Alerstam et al., 2008). Fang et al. recently reported a GPU implementation of the tMCimg

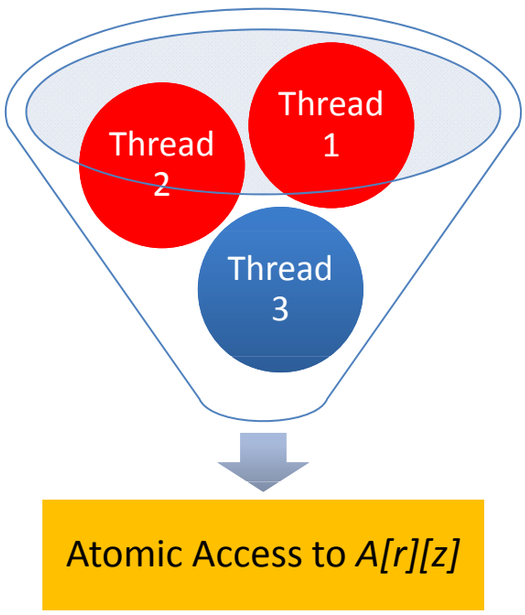


Fig. 5. Concept of an atomic access represented by a funnel: as thread 3 is accessing the absorption array, threads 1 and 2 must wait. Atomic instructions can cause bottlenecks in a computation, especially with thousands of threads common in GPU programming.

code for modelling 3-D voxelized geometry, with a speedup of 300x (without using atomic instructions to ensure data consistency) on the 8800GT graphics card compared to a 1.86GHz Xeon processor (Fang & Boas, 2009). However, the speedup dropped to 75x, or 4 times slower, when atomic instructions were used to guarantee data consistency. Note that one difference between the implementations from these two groups is that Alerstam et al. only used a 1-D vector output (for a time-of-flight histogram with 201 bins), while Fang et al. required a much larger 3-D matrix that needs to be accessed atomically. One could argue that inconsistencies or errors due to non-atomic memory access will only significantly affect the high-fluence region close to the light sources and hence are of little consequence in the critical fluence and threshold models. However, when considering the high photodynamic consumption of oxygen in the high-fluence region, and the resulting PDT-induced hypoxia, a "low dose" region would paradoxically be formed. As a result, for expanded PDT dose distributions, the assumption may not hold true and errors due to non-atomic data accesses can have severe consequences for treatment planning.

This work proposes a different approach to handle the inefficiency in the use of atomic instructions for large 2-D and 3-D result matrices, and addresses the question of how various optimizations can dramatically affect the performance of MC-based simulations for photon migration on NVIDIA GPUs. The final, optimized implementation was also extended to support multiple GPUs to show the possibility of using a cluster of GPUs for complex inverse problems which may require additional computational resources.

4.3 GPU-accelerated MCML Code

In this section, the implementation details of the GPU-accelerated MCML program (named GPU-MCML) are presented, showing how a high level of parallelism is achieved, while avoiding memory bottlenecks caused by atomic instructions and global memory accesses. The optimization process is described to summarize the challenges encountered before arriving at the final solution. This may assist other investigators in related efforts since the MC method is widely applied in computational biophysics and most MC simulations share a set of common features.

4.3.1 Implementation Overview

One difference between writing CUDA code and writing a traditional C program (for sequential execution on a CPU) is the need to devise an efficient parallelization scheme for the case of CUDA programming. Although the syntax used by CUDA is in theory very similar to C, the programming approach differs significantly. Figure 6 shows an overview of the parallelization scheme used to accelerate the MCML code on the NVIDIA GPU. Compared to serial execution on a single CPU where only one photon packet is simulated at a time, the GPU-accelerated version can simulate many photon packets in parallel using multiple threads executed across many scalar processors. Note that the total number of photon packets to be simulated are split equally among all created threads.

The GPU program or kernel contains the computationally intensive part or the key loop in the MCML simulation (represented by the position update, direction update, and fluence update loop in the figure). Other miscellaneous tasks, such as reading the simulation input file, are performed on the host CPU. Each thread executes a similar sequence of instructions, except for different photon packets simulated based on a different random number sequence.

In the current implementation, the kernel configuration is specified as 30 thread blocks (Q=30), each containing 256 threads (P=256). As shown in Figure 6, each thread block is physically mapped onto one of the 30 multiprocessors and the 256 threads interleave their execution on the 8 scalar processors within each multiprocessor. Increasing the number of threads helps to hide the global memory access latency. However, this also increases competition for atomic access to the common $A[r][z]$ array. Therefore, the maximum number of threads, which is 512 threads per thread block on the graphics cards used in this work, was not chosen. A lower number would not be desirable since more than 192 threads are required to avoid delays in accessing a register (due to potential register read-after-write dependencies and register memory bank conflicts (*CUDA Programming Guide 2.3*, 2009)). A similar reasoning applies to the number of thread blocks chosen. A lower number than 30 thread blocks would underutilize the GPU computing resources since there are 30 multiprocessors available. A larger number, such as 60 thread blocks, would decrease the amount of shared memory available for caching and also increase competition for access to the $A[r][z]$ array. The need to alleviate the competition for atomic access is discussed in detail next.

4.3.2 Key Performance Bottleneck

To understand further why atomic accesses to the $A[r][z]$ array could become a key performance bottleneck, notice that all threads add to the same absorption array in the global memory during the fluence update step. In CUDA, atomic addition is performed using the `atomicAdd` instruction. However, using `atomicAdd` instructions to access the global memory is particularly slow, both because global memory access is a few orders of magnitude slower than that of on-chip memories and because atomicity prevents parallel execution of

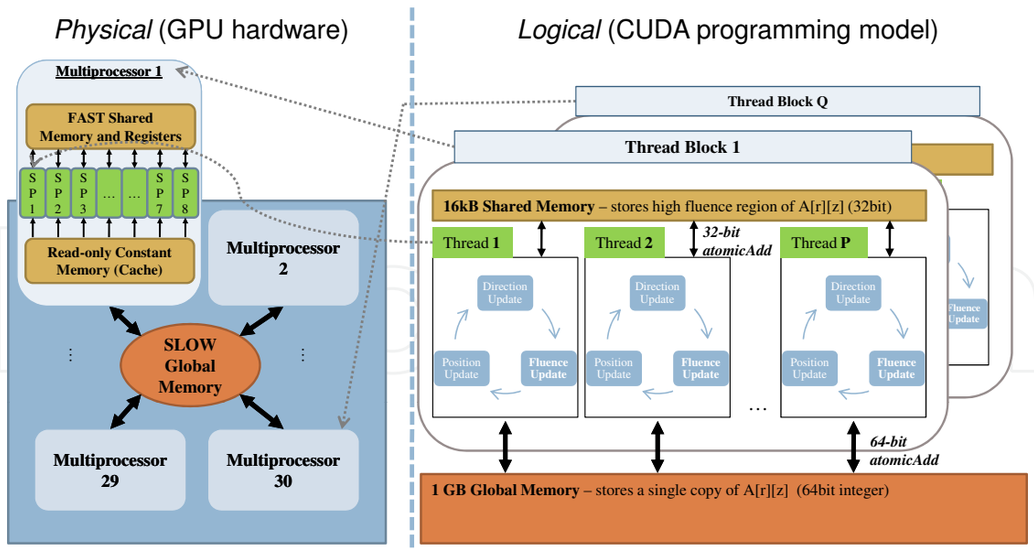


Fig. 6. Parallelization scheme of the GPU-accelerated MCML code (Q=30 and P=256 for each GPU). Note the mapping of the threads to the GPU hardware. In general, this is a many-to-one mapping.

the code (by stalling other threads in the code segment where atomic instructions are located). This worsens with increasing number of threads due to the higher probability of simultaneous access to an element, also known as contention.

Note that although the $A[r][z]$ array could, in theory, be replicated per thread to completely avoid atomic instructions, this approach is limited by the size of the device memory and would not be feasible in the general 3-D case with much larger absorption arrays. Therefore, a more general approach was explored to solve this performance problem.

4.3.3 Solution to Performance Issue

To reduce contention and access time to the $A[r][z]$ array, two memory optimizations, caching in registers and shared memory, were applied.

The first optimization is based on the idea of storing the recent write history, representing past absorption events, in temporary registers to reduce the number of atomic accesses to the global memory. It was observed that consecutive absorption events can happen at nearby, or sometimes the same, locations in the $A[r][z]$ array, depending on the absorption grid geometry and optical properties of the layers. Since the number of registers is limited, in the final solution, only the most recent write history is stored in 2 registers – one for the last memory location and one for the total accumulated weight. In each thread, consecutive writes to the same location of the $A[r][z]$ array are accumulated in these registers until a different memory location is computed. Once a different location is detected, the total accumulated weight in the temporary register is flushed into the global memory using an `atomicAdd` operation and the whole process is repeated.

The second optimization, illustrated in Figure 6, is based on the high event rate, and hence memory access rate, for the $A[r][z]$ elements near the photon source (or at the origin in the MCML model), causing significant contention when atomic instructions are used. Therefore, the region of the $A[r][z]$ array near the source is cached in the shared memory. This optimization has two significant implications. First of all, contention in the most competitive region

of the $A[r][z]$ array is reduced by up to 30-fold since the shared memory copy of the array is updated atomically by only 256 threads within each thread block instead of 7680 threads across 30 blocks. Second of all, accesses to the shared memory are ∼100-fold faster than those to the global memory. Together, these two factors explain the significant improvement in performance (∼2x) observed after this optimization. (Note that the 3000-fold improvement suggested earlier is an optimistic upper bound estimate and is not likely attainable due to the small size of the shared memory and other technical limitations such as shared memory bank conflicts.)

To store as many elements near the photon source as possible in the shared memory, the size of each element in the $A[r][z]$ array was reduced to 32 bits (as opposed to 64 bits for the master copy in the global memory). Given the size of the shared memory is 16 kB, 3584 32-bit elements can be cached compared to only 1792 elements if 64-bit elements were used (3584 x 32 bits or 4 bytes = 14 kB, with the remaining shared memory space allocated elsewhere). However, this reduction also causes a greater risk of computational overflow, which occurs when the accumulated value exceeds $\sim 2^{32}$ (instead of $\sim 2^{64}$ in the 64-bit case). To prevent overflow, the old value is always checked before adding a new value. If overflow is imminent, the value is flushed to the absorption array in global memory, which still uses a 64-bit integer representation. From this calculation, it also becomes evident that 32-bit shared memory entries may not be optimal for 3D applications and 16 bits may be preferable to better cover the larger high access volume.

As an additional optimization technique to avoid atomic access, in the GPU version, photon packets at locations beyond the coverage of the absorption grid no longer accumulate their weights at the perimeter of the grid, unlike in the original MCML code. Note that these boundary elements were known to give invalid values in the original MCML code (Wang et al., 1995). This optimization does not change the correctness of the simulation, yet it ensures that performance is not degraded if the size of the detection grid is decreased, which forces photon packets to be absorbed at boundary elements (significantly increasing contention and access latency to these elements in the $A[r][z]$ array).

4.3.4 Other Key Optimizations

Another major problem with the original MCML code for GPU-based implementation is its abundance of branches (e.g., `if` statements), leading to significant code divergence. In the CUDA implementation, the function for computing the internal reflectance and determining whether a photon packet is transmitted or reflected at a tissue interface was significantly restructured to remove or to reduce the size of a large number of branches.

Finally, this implementation also includes a number of other optimizations, such as using GPU-intrinsic math functions (namely __sincosf(x) and __logf(x)), reducing local memory usage by expanding arrays into individual elements, and storing read-only tissue layer specifications in constant memory.

4.3.5 Scaling to Multiple GPUs

To scale the single-GPU implementation to multiple GPUs, multiple host threads were created on the CPU side to simultaneously launch multiple kernels, to coordinate data transfer to and from each GPU, and to sum up the partial results generated by the GPUs for final output. The same kernel and associated kernel configuration were replicated N times where N is the number of GPUs, except that each GPU initializes a different set of seeds for the random number generator and declares a separate absorption array. This allows the independent

simulation of photon packets on multiple GPUs, similar to the approach taken in CPU-based cluster computing.

4.4 Performance

The execution time of the GPU-accelerated MCML program (named GPU-MCML) was first measured on a single GPU — the NVIDIA GTX 280 graphics card — with 30 multiprocessors. The code was migrated to a Quad-GPU system consisting of two NVIDIA GTX 280 graphics cards and a NVIDIA GTX 295 graphics card with 2 GPUs. This Quad-GPU system contains a total of 120 multiprocessors. The final GPU-MCML was compiled using the CUDA Toolkit and was tested in both a Linux and Windows environment. The number of GPUs used can be varied at run-time and the simulation is split equally among the specified number of GPUs. For baseline performance comparison, a high-performance Intel Xeon processor (Xeon 5160) was selected. The original, CPU-based MCML program (named here CPU-MCML) was compiled with the highest optimization level (gcc -O3 flag) and its execution time was measured on one of the two available cores on the Intel processor.

4.4.1 Skin Model

For performance comparison, a seven-layer skin model at λ=600 nm (shown in Table 1) (Meglinsky & Matcher, 2001) was used. Table 2 shows the execution time of the GPU-MCML program as the number of GPUs was increased. In all cases, the kernel configuration for each GPU was fixed at 30 thread blocks, each with 256 threads. Using one GTX 280 graphics card or 1 GPU with 30 multiprocessors (which contain a total of 240 scalar processors), the speedup was 309x when absorption, reflectance, and transmittance were recorded. The speedup increased to 483 x when absorption was not recorded. Using all 4 GPUs or equivalently 960 scalar processors, the simulation time for 50 million photon packets in the skin model was reduced from approximately 3 h on an Intel processor to only 9.7 s on 4 GPUs. This represents an overall speedup of 1101x ! When only reflectance and transmittance were recorded, the simulation took 5.9 s (1810x) ! Note that the overhead of synchronization between the GPUs and the summation of the partial simulation results would not be noticeable with larger simulation runs.

Layer	n	μ_a (cm^{-1})	μ_s (cm^{-1})	g	Thickness (cm)
1. stratum corneum	1.53	0.2	1000	0.9	0.002
2. living epidermis	1.34	0.15	400	0.85	0.008
3. papillary dermis	1.4	0.7	300	0.8	0.01
4. upper blood net dermis	1.39	1	350	0.9	0.008
5. dermis	1.4	0.7	200	0.76	0.162
6. deep blood net dermis	1.39	1	350	0.95	0.02
7. subcutaneous fat	1.44	0.3	150	0.8	0.59

Table 1. Tissue optical properties of a seven-layer skin model (λ=600 nm).

4.5 Validation

Figure 7 shows the simulated fluence distribution after launching 10^7 photon packets in the skin model shown in Table 1. The outputs produced by the GPU-MCML and CPU-MCML programs match very well. To further quantify any potential error introduced in the imple-

Number of GPUs	Platform Configuration	Time (s)	Speedup
1	GTX 280	34.6 (22.1)	309x (483x)
2	2 x GTX 280	17.5 (11.3)	610x (945x)
3	1 x GTX 280 + GTX 295 (2 GPUs)	12.7 (7.6)	841x (1405x)
4	2 x GTX 280 + GTX 295 (2 GPUs)	9.7 (5.9)	1101x (1810x)

Table 2. Speedup as a function of the number of GPUs for simulating 5x10^7 photon packets in a skin model (λ=600 nm). Baseline (1x) execution time on the Intel Xeon CPU was 10680 s or $\sim$3 h. Values in brackets were generated without tracking absorption; only reflectance and transmittance were recorded.

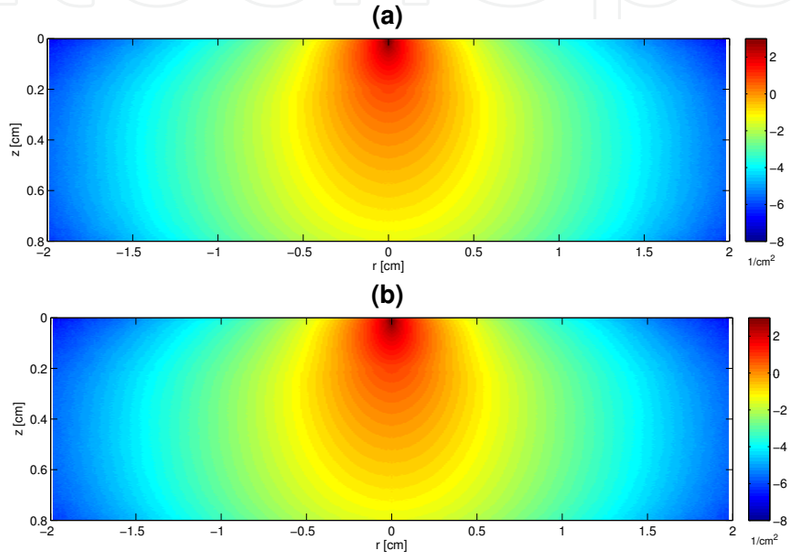


Fig. 7. Logarithm of simulated fluence distribution in the skin model (10^7 photon packets) for the impulse response: (a) generated by GPU-MCML, (b) generated by CPU-MCML.

mentation, the relative error $E[i_r][i_z]$ is computed for each voxel using Equation 11.

$$E[i_r][i_z] = \frac{|A_{gpu}[i_r][i_z] - A_{cpu}[i_r][i_z]|}{A_{cpu}[i_r][i_z]} \tag{11}$$

where A_{cpu} is the gold standard absorption array produced by the CPU-MCML software while A_{gpu} contains the corresponding elements produced by the GPU-MCML program. Figure 8 plots the relative error as a function of position, showing that the differences observed are within the statistical uncertainties between two simulation runs of the gold standard CPU-MCML program using the same number of photon packets.

5. Current challenges

The current challenges in this field mainly arise from the complexity of light-tissue interaction and that of the implementation of full 3D models in hardware. The complicated interactions among light treatment parameters - namely the light fluence rate, photosensitizer, and

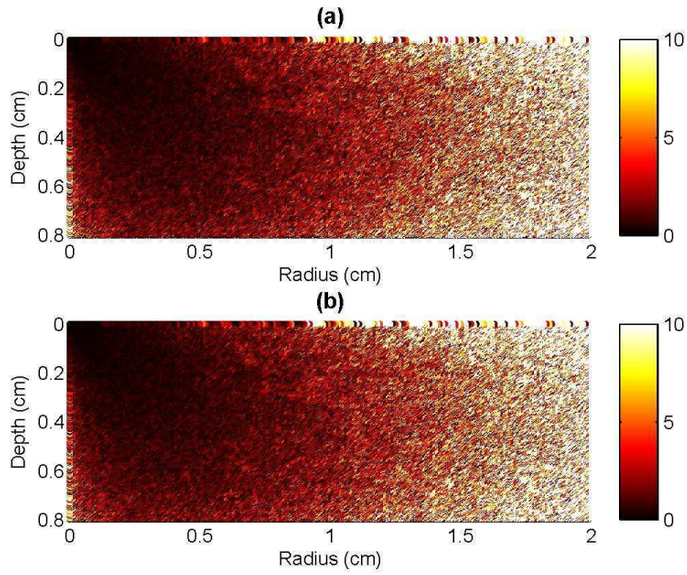


Fig. 8. Distribution of relative error for the skin model (10^7 photon packets): (a) GPU-MCML vs. CPU-MCML, (b) CPU-MCML vs. CPU-MCML. Colour bar represents percent error from 0% to 10%.

ground-state oxygen for IPDT or light absorption and bioheat equation for ILH and ILP - clearly indicate that a solely light-based approach to treatment planning and monitoring of delivered dose will be insufficient in attaining the best achievable clinical outcome, particularly in cases of tissue hypoxia, heterogeneity in photosensitizer concentration, heterogeneity in tissue optical properties, or presence of major vessels. In ILP and ILH, major vessels will result in significant local heat convection. Conversely, the characteristic leakiness of the tumour blood vessels plays a part in the preferential accumulation of PS in the tumour, and their characteristic tortuosity leads to hypoxic or anoxic regions, resulting in poor PDT efficacy. Additionally, over the course of an IPDT treatment, there will be events such as photosensitizer photobleaching, vascular shutdown, oxygen depletion, or inflammation, which will affect PDT efficacy. These complications are driving technical advances in monitoring one or more of the three efficacy-determining parameters: fluence rate, concentration of photosensitizer, and tissue pO_2. Ideally, treatment monitoring devices would capture the dynamics of the photochemical reactions and the spatial heterogeneities as dictated by the anatomy and physiology of the target. The desired spatial sampling of the light fluence rate will be on the order of μ_{eff}, approximately 4 cm^{-1}. For PDT, temporally, a relatively low sampling rate, around 0.03 Hz, is sufficient to capture the changes in tissue optical properties due to vascular events such as inflammation or thrombus formation which occur on the order of minutes. For ILP in particular, a sampling rate closer to 1 Hz may be required. The spatial distribution of the photosensitizer will depend on its delivery via the vasculature; for an intercapillary distance of 100 μm, one would want at least 2 samples per 100 μm, or 0.02 μm^{-1}. Photobleaching will affect the temporal variation in [PS]; thus, sampling at 0.05 Hz is desired, based on reported rates of photobleaching in vitro (Kruijt et al., 2009). No such monitoring is required for ILP and ILH. As for the photosensitizer, availability of molecular oxygen is dependent on

the vasculature - spatially 0.02 μm^{-1} is the approximate sampling resolution goal. For a PDT consumption rate of 30 μMs^{-1}, a temporal sampling rate of 0.08 Hz is required.
While it is possible to monitor these quantities at selected points using implanted optical sensors, there is a limit on the number of fibres which can be inserted and thus a limit on the spatial resolution of treatment monitoring; the volume will be under-sampled, and means to extrapolate the desired quantity to the entire volume are required. These dynamic changes in the CTV and the OAR need to be considered and the treatment plan should be re-adjusted in real time, as discussed further in Section 6. In terms of the hardware implementation of the complete computational framework, memory access time is currently an important consideration for real-time treatment planning. This problem is exacerbated in 3D treatment planning, and it needs to be addressed.

6. Future direction: Real-time, adaptive treatment planning

With the rapid improvement of GPU hardware and the release of the next-generation Fermi GPU architecture for general-purpose computing, GPU-based, real-time treatment planning may soon become a reality. In particular, the new Fermi GPU architecture from NVIDIA features a new memory/cache architecture that offers better memory access time, including faster atomic accesses which are especially important for 3D treatment planning. The dramatic reduction in treatment planning time potentially accomplished by a GPU cluster may, in the future, enable real-time adaptive treatment planning based on the most recent dose parameters obtained from the treatment volume. Currently, pretreatment models assume constant values for tissue optical properties based on population-averaged historical data and ignore the dynamic nature of tissues over the course of the therapy, which directly affects treatment outcomes in interstitial light therapies, especially for ILH and ILP. The implications of real-time dosimetry on the parameter space for optimization are also important. For example, the post-implantation constraints in the optical fibre positions would result in a more confined search space, making simulated annealing an even more attractive approach. From the original N-dimensional search space, only the total power per optical source fibre would remain. However, time-dependent changes in light-tissue interaction parameters and treatment efficacy determining coefficients require frequent execution of the algorithm. Considering a typical PDT treatment lasts 10 to 60 minutes, a temporal resolution of $\sim$5 seconds can be set for real-time computation as an initial research goal. Finally, to realize the full potential of real-time treatment planning, there is a need for more comprehensive dosimetry models that take into account not only physical parameters, but also biological parameters, possibly through real-time treatment monitoring. As we move towards conformal interstitial light therapies with the development of a real-time computational framework for treatment planning, (Lo, W.C.Y. et al., 2010) simulated annealing will likely become an indispensable tool for exploring the increasingly sophisticated landscape of optimization.

7. Acknowledgements

The authors wish to acknowledge the funding support from NSERC and CIHR as well as the contributions of David Han and Erik Alerstam to code development. Research infrastructure support was provided by the Ontario Ministry of Health and Long Term Care (OMHLTC). The views expressed do not necessarily reflect those of OMHLTC.

The GPU implementation described in this chapter has been integrated with the CUDAMCML software (Alerstam et al.). The most updated source code and documentation can be downloaded from ***http://code.google.com/p/gpumcml/***

8. References

Alerstam, E., Svensson, T. & Andersson-Engels, S. (2008). Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration, *Journal of Biomedical Optics* **13**: 060504.

Altschuler, M., Zhu, T., Li, J. & Hahn, S. (2005). Optimized interstitial PDT prostate treatment planning with the Cimmino feasibility algorithm, *Medical Physics* **32**: 3524.

Aniola, J., Selman, S., Lilge, L., Keck, R. & Jankun, J. (2003). Spatial distribution of liposome encapsulated tin etiopurpurin dichloride (SnET2) in the canine prostate: Implications for computer simulation of photodynamic therapy, *International Journal of Molecular Medicine* **11**: 287–292.

ASAP - Getting Started Guide (2009). *Breault Research Organization* . `http://www.breault.com/resources/kbasePDF/broman0108_getstart.pdf`.

Aubry, J., Beaulieu, F., Sévigny, C., Beaulieu, L. & Tremblay, D. (2006). Multiobjective optimization with a modified simulated annealing algorithm for external beam radiotherapy treatment planning, *Medical Physics* **33**: 4718.

Beaulieu, F., Beaulieu, L., Tremblay, D. & Roy, R. (2004). Simultaneous optimization of beam orientations, wedge filters and field weights for inverse planning with anatomy-based MLC fields, *Medical Physics* **31**: 1546.

Bortfeld, T. (2006). IMRT: a review and preview, *Physics in medicine and biology* **51**: R363.

Cheong, W., Prahl, S. & Welch, A. (1990). A review of the optical properties of biological tissues, *IEEE Journal of Quantum Electronics* **26**(12): 2166–2185.

CUDA Programming Guide 2.3 (2009). *NVIDIA Corporation* . `http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.3.pdf`.

Davidson, S., Weersink, R., Haider, M., Gertner, M., Bogaards, A., Giewercer, D., Scherz, A., Sherar, M., Elhilali, M., Chin, J. et al. (2009). Treatment planning and dose analysis for interstitial photodynamic therapy of prostate cancer, *Physics in Medicine and Biology* **54**(8): 2293–2313.

Di Paolo, A. & Bocci, G. (2007). Drug distribution in tumors: mechanisms, role in drug resistance, and methods for modification, *Current Oncology Reports* **9**(2): 109–114.

Fang, Q. & Boas, D. A. (2009). Monte carlo simulation of photon migration in 3d turbid media accelerated by graphics processing units, *Opt. Express* **17**(22): 20178–20190.

Farrell, T., Hawkes, R., Patterson, M. & Wilson, B. (1998). Modeling of photosensitizer fluorescence emission and photobleaching for photodynamic therapy dosimetry, *Applied Optics* **37**: 7168–7183.

Goldstine, H. & Goldstine, A. (1996). The electronic numerical integrator and computer (ENIAC), *IEEE Annals of the History of Computing* pp. 10–16.

Henriques Jr, F. & Moritz, A. (1947). Studies of Thermal Injury: I. The Conduction of Heat to and through Skin and the Temperatures Attained Therein. A Theoretical and an Experimental Investigation*, *The American Journal of Pathology* **23**(4): 530.

Ishimaru, A. (1977). Theory and application of wave propagation and scattering in random media, *Proceedings of the IEEE* **65**(7): 1030–1061.

Jacques, S. & Pogue, B. (2008). Tutorial on diffuse light transport, *Journal of Biomedical Optics* **13**: 041302.

Jankun, J., Lilge, L., Douplik, A., Keck, R., Pestka, M., Szkudlarek, M., Stevens, P., Lee, R. & Selman, S. (2004). Optical characteristics of the canine prostate at 665 nm sensitized with tin etiopurpurin dichloride: need for real-time monitoring of photodynamic therapy, *The Journal of urology* **172**(2): 739–743.

Johansson, A., Axelsson, J., Andersson-Engels, S. & Swartling, J. (2007). Realtime light dosimetry software tools for interstitial photodynamic therapy of the human prostate, *Medical Physics* **34**: 4309.

Kahn, H. & Marshall, A. (1953). Methods of reducing sample size in Monte Carlo computations, *Journal of the Operations Research Society of America* pp. 263–278.

Kruijt, B. et al. (2009). Monitoring interstitial m-THPC-PDT in vivo using fluorescence and reflectance spectroscopy, *Lasers in Surgery and Medicine* **41**(9): 653–664.

Lessard, E. & Pouliot, J. (2001). Inverse planning anatomy-based dose optimization for HDR-brachytherapy of the prostate using fast simulated annealing algorithm and dedicated objective function, *Medical Physics* **28**: 773.

Ma, C., Mok, E., Kapur, A., Pawlicki, T., Findley, D., Brain, S., Forster, K. & Boyer, A. (1999). Clinical implementation of a Monte Carlo treatment planning system, *Medical Physics* **26**: 2133.

Martin, A., Roy, J., Beaulieu, L., Pouliot, J., Harel, F. & Vigneault, E. (2007). Permanent prostate implant using high activity seeds and inverse planning with fast simulated annealing algorithm: A 12-year Canadian experience, *International Journal of Radiation Oncology Biology Physics* **67**(2): 334–341.

Meglinsky, I. & Matcher, S. (2001). Modelling the sampling volume for skin blood oxygenation measurements, *Medical and Biological Engineering and Computing* **39**(1): 44–50.

Metropolis, N. (1989). The beginning of the Monte Carlo method, *From Cardinals to Chaos: Reflections on the Life and Legacy of Stanislaw Ulam* p. 125.

Metropolis, N. & Ulam, S. (1949). The monte carlo method, *Journal of the American Statistical Association* pp. 335–341.

Morrill, S., Lam, K., Lane, R., Langer, M. & Rosen, I. (1995). Very fast simulated reannealing in radiation therapy treatment plan optimization, *International Journal of Radiation Oncology Biology Physics* **31**: 179–179.

Niedre, M., Secord, A., Patterson, M. & Wilson, B. (2003). In vitro tests of the validity of singlet oxygen luminescence measurements as a dose metric in photodynamic therapy, *Cancer Research* **63**(22): 7986.

Pang, A., Smith, A., Nuin, P. & Tillier, E. (2005). SIMPROT: using an empirically determined indel distribution in simulations of protein evolution, *BMC bioinformatics* **6**(1): 236.

Pennes, H. (1948). Analysis of tissue and arterial blood temperatures in the resting human forearm, *Journal of Applied Physiology* **1**(2): 93.

Pogue, M. (1994). Mathematical model for time-resolved and frequency-domain fluorescence spectroscopy in biological tissues, *Appl. Opt* **33**: 1963–1974.

Rendon, A. (2008). *Biological and Physical Strategies to Improve the Therapeutic Index of Photodynamic Therapy*, PhD thesis, University of Toronto.

Rivard, M., Coursey, B., DeWerd, L., Hanson, W., Huq, M., Ibbott, G., Mitch, M., Nath, R. & Williamson, J. (2004). Update of AAPM Task Group No. 43 Report: A revised AAPM protocol for brachytherapy dose calculations, *Medical Physics* **31**: 633.

Scheidler, J., Hricak, H., Vigneron, D., Yu, K., Sokolov, D., Huang, L., Zaloudek, C., Nelson, S., Carroll, P. & Kurhanewicz, J. (1999). Prostate cancer: localization with three-dimensional proton MR spectroscopic imaging-clinicopathologic study, *Radiology* **213**(2): 473.

Tuchin, V. (1997). Light scattering study of tissues, *Physics-Uspekhi* **40**(5): 495–515.

Wang, K., Mitra, S. & Foster, T. (2007). A comprehensive mathematical model of microscopic dose deposition in photodynamic therapy, *Medical Physics* **34**: 282.

Wang, L., Jacques, S. & Zheng, L. (1995). MCML - Monte Carlo modeling of light transport in multi-layered tissues, *Computer Methods and Programs in Biomedicine* **47**(2): 131–146.

Wang, L., Jacques, S. & Zheng, L. (1997). CONV - convolution for responses to a finite diameter photon beam incident on multi-layered tissues, *Computer Methods and Programs in Biomedicine* **54**(3): 141–150.

Weishaupt, K., Gomer, C. & Dougherty, T. (1976). Identification of singlet oxygen as the cytotoxic agent in photo-inactivation of a murine tumor, *Cancer Research* **36**(7 Part 1): 2326.

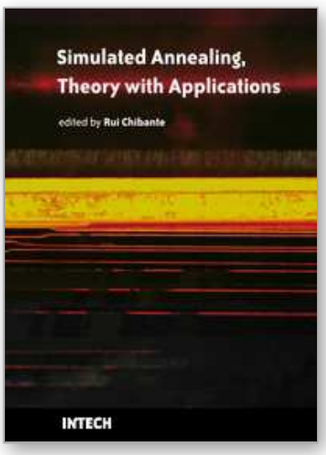
Zhu, T., Finlay, J., Zhou, X. & Li, J. (2007). Macroscopic modeling of the singlet oxygen production during PDT, *Proceedings of SPIE*, Vol. 6427, p. 642708.

FPGA-based Monte Carlo computation of light absorption for photodynamic cancer therapy,2009 17th IEEE Symposium on Field Programmable Custom Computing Machines, 157–164, IEEE.

Hardware acceleration of a Monte Carlo simulation for photodynamic therapy treatment planning,Journal of Biomedical Optics, Vol. 14, p. 014019.

GPU-accelerated Monte Carlo simulation for photodynamic therapy treatment planning,Proceedings of SPIE, Vol. 7373, p. 737313.

Computational Acceleration for Medical Treatment Planning: Monte Carlo Simulation of Light Therapies Accelerated using GPUs and FPGAs,VDM Verlag Dr. Muller, ISBN: 978-3639250381.



Simulated Annealing, Theory with Applications
Edited by Rui Chibante

ISBN 978-953-307-134-3
Hard cover, 292 pages
Publisher Sciyo
Published online 18, August, 2010
Published in print edition August, 2010

The book contains 15 chapters presenting recent contributions of top researchers working with Simulated Annealing (SA). Although it represents a small sample of the research activity on SA, the book will certainly serve as a valuable tool for researchers interested in getting involved in this multidisciplinary field. In fact, one of the salient features is that the book is highly multidisciplinary in terms of application areas since it assembles experts from the fields of Biology, Telecommunications, Geology, Electronics and Medicine.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen